DOCUMENT RESUME

ED 344 492 FL 020 356

AUTHOR Breland, Hunter M.; Jones, Robert J.
TITLE A Lexical Universe for Verbal Reasoning Tests.
PUB DATE Apr 92
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS College Freshmen; Comparative Analysis; *Computational Linguistics; High Schools; High School Students; Language Tests; Mathematical Linguistics; *Statistical Analysis; Test Construction; *Thinking Skills; *Verbal Tests; *Vocabulary; *Word Frequency; Word Recognition
IDENTIFIERS *Reasoning Tests

ABSTRACT

The usefulness of a "lexical universe" in the development of verbal reasoning tests was studied. The lexical universe, which is a comprehensive listing of words and their frequencies based on a collection of machine-readable texts exceeding 14 million words of running text, is an outcome of a 1987 College Board study of the language in materials encountered by high school students and first-year college students. This report describes an analysis of the materials collected and presents statistics on the following: total frequencies; an index of dispersion across text categories; an estimate of frequency per one million tokens; and a standard frequency index based on a logarithmic transformation. Additional tables show some sample words and statistics for the College Board corpus as compared to the American Heritage corpus, and a comparison of frequency statistics for a selection of 20 words based on 4 different corpora. The difficulty of acquiring and maintaining a large vocabulary is discussed, along with the importance of word frequency and word frequency counts. It is concluded that the major advantage of word frequency counts is that they yield difficulty estimates for all words that are encountered in the corpus. Contains 24 references. (LB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED344492

A Lexical Universe for Verbal Reasoning Tests

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Hunter Breland
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Hunter M. Breland and Robert J. Jones

Educational Testing Service

Paper prepared for presentation at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April 1992.

FL 020 356

BEST COPY AVAILABLE

A Lexical Universe for Verbal Reasoning Tests

Hunter M. Breland and Robert J. Jones

Educational Testing Service

In 1987, a study was initiated by the College Board to obtain a better understanding of the language in materials that high school students and first-year college students are exposed to. Previous investigations of this type were considered out of date, based on too small of a corpus, or based on samples of text inappropriate for the high school to college transition. The well-known Thorndike & Lorge (1944) study was based on samples collected in the 1920's and 1930's. The Brown corpus (Kucera & Francis, 1969; Francis & Kucera, 1982) was based on only one million words of text. The American Heritage word frequency count (Carroll, Davies, and Richman, 1971) was limited to reading materials used in grades three through nine. The objective of the College Board study was to obtain an up-to-date collection of reading materials used in high school and the first year of college. One of the outcomes of ths study is a comprehensive listing of words and their frequencies based on over 14 million words of running text. This listing constitutes a lexical universe that is quite useful in the development of verbal reasoning tests.

The first task was to make a listing of reading materials from existing curriculum surveys, curriculum guides, reading lists, and other sources indicating the degree to which specific works, textbooks, periodicals, and other materials were in use in 1987. Second, a ranking of the importance of specific reading materials was conducted based upon the frequency with which these materials were cited (see Tables 1, 2, & 3). Women and minority writers were included who did not rank high by this procedure, however. Third, a survey of materials available in electronic form was conducted. Finally, materials were selected that were both of importance and available in machine-

readable form. The result was a collection of machine-readable texts exceeding 14 million words of running text (see Table 4). The materials collected were analyzed using the WORDCRUNCHER text indexing and retrieval software, and word frequency files generated were uploaded to the mainframe and combined into a single file. Statistical indexes were computed following the methods of Carroll et al. (1971). These statistics included total frequencies (F), an index of dispersion across text categories (D), an estimate of frequency per one million tokens (U), and a standard frequency index (SFI) based on a logarithmic transformation (see Table 5). Table 6 shows some sample words and statistics for the College Board (CB) corpus as compared to the American Heritage (AH) corpus (Carroll et al., 1971).

Table 7 presents a comparison of frequency statistics for a selection of 20 words based on four different corpora. Some of these words were chosen because of probable changes in usage over time. The Thorndike & Lorge (TL) corpus is the oldest (about 1930), the American Heritage corpus was collected in about 1969, the Kucera-Francis (KF) corpus in about 1966, and the College Board corpus in 1987-88. It is not surprising that words like environment, electron, organic, ethnic, ecology, racism, and telecommunications occur more frequently in the later collections. These differences highlight the need for frequent updates of word frequency statistics. It is also of interest to note that some of the words in the list occurred infrequently in the smaller corpora. Seven of the 20 words did not occur at all (U = 0), and three words occurred only once (hapax legomena) in the AH and KF corpora. Such nonoccurrences highlight the importance of corpus size for word frequency studies. Any given word may not occur at all either because it is not in current usage or because the corpus is not large enough. Even with corpora of relatively large size, as for the TL and CB corpora, the statistics

computed for relatively infrequent words have considerable error.

The lexical universe constructed represents 14 million words of text of the type that high school and college students are likely to encounter. A student who begins reading at age 6 would have read this many words by age 17 if he or she read 1.27 million words a year (3,479 words per day). Nagy, Anderson, & Herman (1987) estimate that the typical fifth-grade student reads about a million words per year. Thus, the universe is representative of the amount of reading that a typical student would complete during the first 14 years of reading and the amount more prolific readers would complete during fewer years of reading.

It has been estimated that the vocabulary size for first-graders averages about 5,000 words and that for college students about 50,000 (Lorge & Chall, 1963). These estimates indicate the very rapid growth of vocabulary during the school years. Studies show, however, that there exist very large individual differences in vocabulary acquisition. Seashore & Eckerson (1940) found that the top 10 percent of college students had vocabularies twice the size of the lowest 10 percent.

A large vocabulary is difficult to acquire and maintain because many important words occur infrequently in a language. Nagy & Anderson found that only about 3,000 words occurred more than ten times per million words of running text and only about 10,000 words occurred more than once per million words of running text. If a college student is to maintain a vocabulary of 50,000 words, four-fifths of those words will be encountered relatively infrequently in normal reading and thus have to be remembered. Some important words occur quite rarely and, accordingly, the task of acquiring and maintaining an understanding of such words is not as simple as it may seem.

Further evidence of the importance of word frequency comes from

linguistic studies, studies of verbal comprehension, and studies of reading. The assumption that the words most commonly used are learned faster and remembered better is supported by a number of studies (Brown, 1958; Clark, 1983; Howes, 1954; Howes & Solomon, 1951; Solomon & Howes, 1951; Sternberg & Powell, 1983; Tomayo, 1987; Werner & Kaplan, 1952). The speed with which a reader can access a word's meaning depends on how frequently that word has been encountered in the past (Just & Carpenter, 1987). Words that are encountered most frequently in a language are processed with the greatest speed. It has been demonstrated that frequently-occurring words are easier to recognize during a brief presentation (Howes & Solomon, 1951). It has also been shown that frequently-occurring words take less time to classify (Glanzer & Ehrenreich, 1979; Rubenstein, Garfield, & Millikan, 1970). Just & Carpenter (1980) studied the relationship between word frequency and what they termed "gaze-duration," the amount of time a reader spends looking at a word. It was demonstrated that more frequent words had a lower mean gaze duration than did less frequent words. Other researchers have also shown that readers spend less time looking at more frequent words (Gordon, 1985; Rayner, 1983). It is clear from these kinds of studies that the time required to read and understand a word is related to its normative frequency in the language.

Now there are criticisms that can be made of word frequency counts and the use of them as an index of word difficulty. First, it is obvious that different words that are spelled the same but have different meanings depending on the context in which they are used cannot be differentiated in the usual computer analysis. Thus a word like "book," for example, that has an entirely different meaning in a police station than in a library will be treated as only a single word. The police station usage is undoubtedly a more difficult word than is the library usage, however. A second criticism that

can be made is that most counts are never quite large enough to yield accurate estimates of the difficulty of relatively rare words. A third criticism is that a study of written materials ignores spoken language, which is very different. All of these criticisms are legitimate ones.

But consider the alternative to word frequency counts. It is possible to actually test word difficulty by asking students and adults of different ages to respond to multiple-choice questions about them. But there are two major problems with this approach. First, the number of words tested will be limited. Second, the accuracy of the estimates obtained by this approach are a function of the quality of the distractors, and the distractors as well as the key need to be less difficult than the word being studied. How does one know that the options are less difficult than the word being studied? Even so, the distractors may not work well.

Word frequency counts yield difficulty estimates for <u>all</u> words that are encountered in the corpus, and that is their major advantage. If their limitations are recognized, they can still be quite useful. The College Board study results are frequently used by ETS test developers to resolve disputes. As one example that occurred not too long ago, the word "contrail" was questioned by a test reviewer as being arcane. But other test developers disputed this contention, saying that the word was very common. Consultation of the College Board study revealed that the word occurred only two times in the 14 million word sample and, accordingly, was not very common in the kinds of materials that most high school students read. Further research revealed that the test developers who tended to know the word were mostly males.

References

Brown, R. (1958). How shall a thing be called? Psychological Review, 65, 14-21.

Carpenter, P. A., & Daneman, M. (1981). Lexical retrieval and error recovery in reading: A model based on eye fixations. Journal of Verbal Learning and Verbal Behavior, 20, 137-160.

Carroll, J. B., Davies, P., & Richman, B. (1971). The American Heritage Word Frequency Book. New York: American Heritage Publishing Co., Inc.

Carroll, J. B., & White, M. N. (1973). Age-of acquisition norms for 220 picturable nouns. Journal of Verbal Learning and Verbal Behavior, 12, 563-576.

Clark, E. V. (1983). Meanings and concepts. In P. H. Mussen (Ed.), Handbook of child psychology. (Fourth Edition). Volume II, Cognitive development. (J. H. Flavel and E. M. Markman, Editors). New York: Wiley.

Francis, W. N., & Kucera, H. (1982). Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin Co.

Glanzer, M., & Ehrenreich, S. L. (1979). Structure and search of the internal lexicon. Journal of Verbal Learning and Verbal Behavior, 18, 381-398.

Gordon, B. (1985). Subjective frequency and the lexical decision latency function: Implications for mechanisms of lexical access. Journal of Memory and Language, 24, 631-645.

Howes, D. H. (1954). On the interpretation of word frequency as a variable affecting the speed of recognition. Journal of Experimental Psychology, 48, 106-112.

Howes, D. H. & Solomon, R. L. (1951). Visual duration thresholds as a function of word probability. Journal of Experimental Psychology, 41, 401-410.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87, 329-354.

Just, M. A. & Carpenter, P. A. (1987). The psychology of reading and language comprehension. Boston: Allyn and Bacon, Inc.

Kucera, H. & Francis, W. N. (1967). Computational analysis of present-day American English. Providence, R.I.: Brown University Press.

Lorge, I., & Chall. J. (1963). Estimating the size of vocabularies of children and adults: An analysis of methodological issues. Journal of Experimental Education, 32, 147-157.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? Reading Research Quarterly, 19, 304-330.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. American Educational Research Journal, 24, 237-270.

Rayner, K. (Ed.). (1983). Eye movements in reading: Perceptual and language processes. New York: Academic Press.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. Journal of Verbal Learning and Verbal Behavior, 9, 487-494.

Seashore, R. H., & Eckerson, L. D. (1940). The measurement of individual differences in general English vocabularies. Journal of Educational Psychology, 31, 14-38.

Solomon, R. L. and Howes, D. H. (1951). Word frequency, personal values, and visual duration thresholds. Psychological Review, 58, 256-270.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. American Psychologist, 38, 873-893.

Thorndike, E. L., & Lorge. I. (1944). The Teacher's Word Book of 30,000 Words. New York: Bureau of Publications, Teachers College, Columbia University.

Tomayo, J. M. (1987). Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. Educational and Psychological Measurement, 47, 893-902.

Werner, H. and Kaplan, E. (1952). The acquisition of word meanings: A developmental study. Monographs of the Society for Research in Child Development, #51.

Table 1. Leading Authors

Author	Points	Author	Points
Mark Twain**	28	William Wordsworth**	12
Nathaniel Hawthorne**	26	Robert Louis Stevenson*	12
Ernest Hemingway*	25	Leo Tolstoi*	12
Charles Dickens**	24	Anton Chekov	11
William Shakespeare**	24	William Golding*	11
John Steinbeck**	22	Rudyard Kipling	11
F. Scott Fitzgerald**	21	Jack London**	11
Jane Austen**	20	J. R. R. Tolkien*	11
Joseph Conrad**	20	Edith Wharton*	11
George Orwell*	20	Walt Whitman**	11
William Faulkner**	19	James Baldwin*	10
Thomas Hardy**	19	Robert Browning	10
Henrik Ibsen*	19	Emily Dickinson	10
Herman Melville**	19	Ralph Waldo Emerson**	10
Emily Bronte*	18	Anne Frank*	10
Stephen Crane*	18	John Keats*	10
Arthur Miller*	18	John Knowles*	10
J. D. Salinger*	18	Sinclair Lewis*	10
Jonathon Swift*	18	Mary Shelley*	10
Homer*	17	William Blake**	9
Edgar Alan Poe*	17	Daniel Defoe**	9
George Bernard Shaw	17	Herman Hesse**	9
Harper Lee*	16	Franz Kafka*	9
Henry David Thoreau**	16	D. H. Lawrence	9
Fedor Dostoesvki*	15	Alan Paton*	9
Sophocles	15	Erich Maria Remarque	9
T. S. Eliot	14	Harriet Beecher Stowe*	9
Benjamin Franklin**	14	H. G. Wells	9
Henry James**	14	Oscar Wilde*	9
Ray Bradbury*	13	Thomas Wolfe*	9
Charlotte Bronte*	13	Herman Wouk*	9
George Eliot**	13	William Butler Yeats**	9
Ralph Ellison*	13	Samuel Beckett**	8
Robert Frost*	13	Lord Byron	8
Aldous Huxley*	13	Lewis Carroll**	8
James Joyce**	13	James Fenimore Cooper*	8
Thornton Wilder*	13	Daphne Du Maurier**	8
Tennessee Williams	13	Carson McCullers*	8
Richard Wright*	13	James Michener*	8
Willa Cather**	12	Flannery O'Connor*	8
Miguel de Cervantes*	12	Plato	8
Geoffrey Chaucer	12	Antoine de St. Exupery*	8
A. Conan Doyle*	12	Alexandr Solzhenitsyn*	8
John Milton**	12	Alfred Lord Tennyson	8

*Sampled **At least one complete work sampled

Table 2. Leading Works

Work	Points	Work	Points
The Scarlet Letter**	26	Black Boy*	10
Huckleberry Finn**	24	The Invisible Man*	10
The Great Gatsby**	21	My Antonia	10
Lord of the Flies*	18	Native Son*	10
Great Expectations**	17	Nineteen Eighty-four	10
Hamlet**	17	A Portrait of the Artist**	10
To Kill a Mockingbird*	17	Walden Pond**	10
The Grapes of Wrath*	16	For Whom the Bell Tolls*	10
The Odyssey*	16	All Quiet on the Western Front	9
Wuthering Heights*	16	Canterbury Tales	9
Catcher in the Rye	15	A Christmas Carol**	9
The Crucible*	15	Cry, the Beloved Country*	9
Gulliver's Travels*	15	David Copperfield*	9
Julius Caesar**	15	Dr. Jekyll and Mr. Hyde*	9
Of Mice and Men*	15	Don Quixote*	9
The Old Man and the Sea*	15	Ethan Frome*	9
Pride and Prejudice**	15	The Stranger	9
The Red Badge of Courage*	15	The Sun Also Rises*	9
Romeo and Juliet**	15	Uncle Tom's Cabin*	9
Death of a Salesman*	14	Alice in Wonderland**	8
A Doll's House*	14	Autobiography of Franklin**	8
Macbeth**	14	Babbitt*	8
Our Town*	14	The Brothers Karamazov*	8
A Tale of Two Cities*	14	The Call of the Wild**	8
Billy Budd*	13	The Caine Mutiny*	8
Crime and Punishment*	13	A Farewell to Arms*	8
Moby Dick**	13	The Iliad	8
Pygmalion	13	Lord Jim**	8
Tess of the D'Ubervilles**	13	The Mayor of Casterbridge*	8
Brave New World*	12	The Merchant of Venice	8
The Diary of Anne Frank*	12	A Midsummer Night's Dream**	8
Heart of Darkness*	12	Oliver Twist*	8
Oedipus Rex	12	Othello	8
The Pearl*	12	Rebecca**	8
Tom Sawyer**	12	Robinson Crusoe*	8
Jane Eyre*	11	Silas Marner**	8
A Separate Peace*	10	Waiting for Godot**	8
The Bible**	10	War and Peace*	8

*Sampled **Complete work sampled

Table 3. Textbooks, Periodicals, and Other Materials Sampled

Textbook Types
Art
Biology
Business
Chemistry
English Composition
English Literature
Government
History
Physics
Psychology
Science

Periodicals
Atlantic
Christian Science Monitor
Los Angeles Times
Ladies Home Journal
MacNeil/Lehrer Report
Newsweek
The Nation
National Review
The New Republic
The New York Times
People Magazine
Sports Illustrated
Time Magazine
U. S News and World Report
The Washington Post

Other Materials
The Academic American Encyclopedia
King James Bible**
The Magna Carta**
The Mayflower Compact**
The Federalist Papers**
Common Sense**
The Declaration of Independence**
The Articles of Confederation**
The U. S. Constitution**
Washington's First Inaugural Address**
Washington's Farewell Address**
Jefferson's First Inaugural Address**
The Monroe Doctrine**

**Complete text sampled

Table 4. Text by Categories

Category	Words	%
Fiction Categories		
1. American Novels and Short Stories	2,035,067	14
2. British Novels and Short Stories	2,215,932	15
3. Other Novels and Short Stories	114,202	1
4. Poetry	959,861	7
5. Drama	438,020	3
	5,763,082	40
Non-Fiction Categories		
Subject-matter areas:		
6. Art and Architecture	274,205	2
7. Biography and Autobiography	633,753	4
8. Business, Economics, and Agriculture	175,134	1
9. English, Language, and Literature	903,654	6
10. Geography	197,004	1
11. Government, Law, and Military	430,760	3
12. History	471,344	3
13. Medicine and Health	186,660	1
14. Music and Theater	67,706	5
15. Natural Science	791,279	6
16. Religion, Philosophy, and Mythology	177,693	1
17. Social Sciences and Education	253,372	2
18. Sports and Games	254,537	2
	4,817,101	34
Other Non-Fiction:		
19. Essays	866,918	6
20. The Bible	791,336	6
21. Historical Documents	324,274	2
22. Newspapers & TV News	196,747	1
23. News Magazines	351,905	2
24. Political Magazines	613,676	4
25. Literary Magazines	252,230	2
26. Popular Magazines	198,343	1
27. Women's Magazines	185,272	1
	3,780,701	26
TOTAL	14,360,884	100

Table 5. Statistical Computations

Abbreviation	Description
F	Total observed frequency. The number of occurrences of a word type in the corpus.
D	An index of dispersion which can take values from .00 to 1.00. It is based on the degree to which a word type occurred across the 27 categories of text sampled. D is zero when all of the occurrences of the word type are found in a single category. D can be as high as 1.0 if a word type is found in all categories and certain other requirements are met.
U	The frequency of a word type per one million tokens, adjusted for dispersion over the 27 categories. If D = 1.0, U is simply F/14.36 (rounding). If D is less than 1.0, however, U is adjusted downward to the degree that D is less than 1.0. If D = 0, U takes a minimum value reflecting the average weighted probability of the word type over the 27 text categories.
SFI	Standard Frequency Index. SFI = 10(log U + 4).

SFI	Expected Occurrence of a Word Type
90	Once in every ten tokens
80	Once in every one hundred tokens
70	Once in every one thousand tokens
60	Once in every ten thousand tokens
50	Once in every hundred thousand tokens
40	Once in every one million tokens
30	Once in every ten million tokens
20	Once in every hundred million tokens
10	Once in every one billion tokens

Table 6. Sample Words and SFI's

Word	SFI (CB)	SFI (AH)
the	87.8	88.6
and	85.2	84.2
if	73.3	73.9
fact	63.5	62.4
car	58.7	64.8
whale	55.1	56.3
hook	50.4	52.6
faction	48.1	34.8
torpid	42.9	*
glib	40.0	*
lugubrious	35.4	**
anomie	32.6	**
avuncular	32.3	**
hubris	29.3	**
redaction	27.9	**
oxymoron	25.9	**

*hapax legomenon
**no occurrence

Table 7. Comparisons of Selected Words in Different Corpora

Word	U Value			
	TL	CB	AH	KF
Year Published	1944	1991	1971	1967
Tokens (millions)	18	14	5	1
1. environment	13.89	73.40	63.08	53.25
2. electron	1.50	62.19	41.46	39.45
3. organic	8.06	32.87	11.59	40.43
4. immune	5.39	24.10	7.07	8.88
5. ethnic	.28	19.57	2.55	12.82
6. ecology	.00	12.05	5.90	1.97
7. racism	.00	6.76	1.57	.99
8. telecommunications	.00	5.36	.00	.00
9. vial	2.39	3.62	.20	.00
10. zionism	.50	3.62	.59	1.97
11. torpid	1.56	3.62	.20	.99
12. prosaic	2.61	3.55	.59	1.97
13. blithe	7.22	2.92	.59	4.93
14. hegemony	.33	2.65	.20	.99
15. Chicano	.00	1.60	.00	.00
16. byte	.00	1.25	.00	.00
17. lugubrious	.39	1.11	.00	.00
18. paparazzi	.00	.35	.00	.00
19. hubris	.00	.35	.00	.99
20. oxymoron	.00	.21	.00	.00

Symbols and abbreviations:

U Occurrences per one million tokens

TL Thorndike & Lorge (1944), 18 million tokens

CB College Board corpus, 1991, 14 million tokens

AH American Heritage corpus, 1971, 5 million tokens

KF Kucera & Francis (1967), one million tokens